动 物 学 研 究 1993，14 (2)：186—193 ISSN 0254-5853
Zoological Research CN 53-1040 / Q

综述

分子进化树的构建

吕宝忠
(上海市肿瘤研究所 200032)

关键词：进化树，构树方法，重抽样 系统树；分子进化树
Key words: Evolutionary tree, Tree-making methods, Resampling

进化树 (evolutionary tree) 又名系统树 (phylogenetic tree) 已发展成为多学科（包括生命科学中的进化论、遗传学、分类学、分子生物学、生物化学、生物物理学和生态学，又包括数学中的概率统计、图论、计算机科学和群论）交叉形成的一个边缘领域。闻名国际生物学界的美国冷泉港定量生物学会议于 1987 年特辟出"进化树"专栏进行学术讨论，标志着该领域已成为现代生物学的前沿之一，迄今仍很活跃。

分子进化树（以分子数据为依据构建的进化树）不仅精确地反映物种间或群体间在进化过程中发生的极微细的遗传变异（小至一个氨基酸或一个核苷酸差异），而且借助化石提供的大分子类群的分化年代能定量地估计出物种间或群体间的分化年代，这对进化论的研究而言无疑是一场革命。然而分子进化树的构建尚未臻成熟，因而大有发展余地。

当今不同的构树法，大体可归并成三大类：1. 简约法 (parsimony)；2. 距离矩阵法 (distance matrix methods) 3. 最大似然法 (maximum likelihood methods)。鉴于分子数据系抽样所获，可利用重抽样技术 (resampling) ——最重要的有折刀法 (jackknifing) 和自助法 (bootstrapping) ——检验并校正所获数据。本文拟介绍和分析上述诸法，并展望其美好前景。

三类不同的构树法

根据是否指出最终的共同祖先，将进化树分为有根树（指出最终共同祖先）和无根树（不指出最终共同祖先）两大类（图 1）。从图论原理来说，当分析的物种（或群体）有 n 类，并假定它们完全按二歧式（即一个祖种必然分裂为两个子种）分化，则它们可被构建成$(2n-3)!\ /\ [(2^{n-2}\cdot(n-2)!]$ 种不同类型的有根树。对无根树来说，上述物种构建成的进化树数目相当于(n−1)种物种构建成的不同类型的有根树数目 (Nei, 1987)。构树法的目标在于挑出一棵最终树（点估计）或挑出确定置信限范围的一些最终树（区间估计），

本文 1991 年 10 月 25 日收到，1992 年 8 月 5 日修回。

从而描述历史上出现过的进化历程。

一、简约法

基本原则为，构建成的一棵进化树，其核苷酸或氨基酸的替代总数应取最小值。具体操作时，既不考虑匹配后的全同位点（如下例中的第1、5、6和7位点），也不考虑除一个物种外余皆全同的位点（如例中的第8位点），因为它们并不提供物种间分化的信息（Fitch，1977）。现设想一例有5个物种，其同源核酸的匹配如下：

α: G A A A T T G C
β: G A A C T T G T
γ: G C A C T T G T
δ: G C C C T T G T
ε: G C C A T T G T

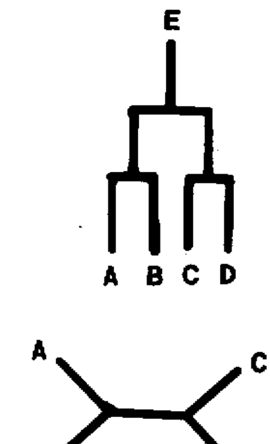


图1 有根树（上）和无根树(下)
Fig. 1 A rooted tree (above) and an unrooted tree (below)

本例中γ与其他物种，几乎都有相等距离，故先暂不考虑它。由于α和β，以及δ和ε都有两次形成单体（所谓单体即两物种在某位如α和β在第2位合并后A在该位仅存单一因子）的机会，故可先将α和β，以及δ和ε分别聚成两类，然后将γ置于上述两聚类群之间，即构成了如图2的进化树。当然，如果匹配的位点中非全同核苷酸较多且物种数目较多时，必须借助计算机程序方能构建分子树。

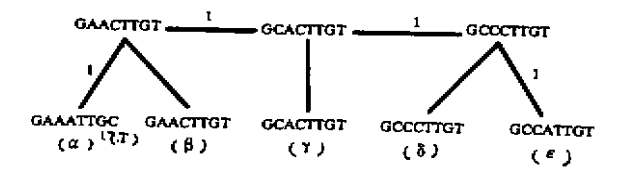


图2 最大简约法构建的分子进化树
(α—ε为现存5物种的部分序列)
Fig. 2 A molecular evolutionary tree reconstructed by the maximum parsimony

图2指出，上述5个物种在整个进化历程中共发生4次替代（物种γ第8位点的C系单体，可能是新发生的，不计在内），这无疑是最小的替代数。

上述的简约法又称为最大简约法，可挑出一棵最终树，还可选出一组树。当选出一组树时，则可根据其他生物学证据，将其中一棵（属于替代总数少的但不一定是最少的，这是因为进化是相当复杂的）树即所谓的生物学树（biological tree）挑出（Beitema等，1986）。其后又有人提出了加权简约法（weighted parsimony）以及进化简约法（evolutionary parsimony）（Lake，1987）。事实上进化简约法属最大似然法范畴，容后介绍。

二、距离矩阵法

此处所谓的距离，既可指相对替代率，也可指遗传距离或进化距离。应该指出，上述两参数构建的进化树不一定有相同意义（吕宝忠等，1992）。常见的距离矩阵法至少有以下几种：

1. UPGMA（unweighted pair-group method with arithmetic mean）（不加权成对群

算术平均法） 先将成对比较的距离排成矩阵（图 3），挑出距离最小者如 D_{34} 作为第 1 个距离期望值（该最小距离期望值等于观察值）。然后把该两物种合并为一个新群以构建新的矩阵。新群与其他两个物种的距离分别为 $D_{1(34)}=(D_{13}+D_{14})/2$ 和 $D_{2(34)}=(D_{23}+D_{24})/2$，找出第二个最小距离。这样逐次归并直到整个物种合并成一个大群，便完成了构树。UPGMA 构建的树既有拓扑（topology，branch pattern）和分枝长度（branch length），还有内结的方差（Nei，1987）。该法有广泛的应用，描述近缘物种间或我国不同民族间亲缘关系和分化演变的进化树大抵应用本法。

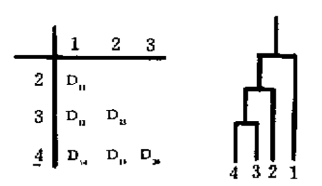


图 3 距离矩阵法(左为 4 物种的距离矩阵，右为由 UPGMA 构建的分子进化树)

Fig. 3 A distance matrix method (A distance matrix of 4 species in left; a molecular evolutionary tree reconstructed by this method in right)

2. Fitch–Margoliash 法（Fitch 与 Margoliash，1967） 如果在 n 个物种中，物种 A 和 B 间的距离最小，则把其他物种组成 C 群，并计算它与 A 和 B 聚类后的共同祖先 D 至 A、B 的分枝长度（图 4）x 和 y，

$$x=(d_{AB}+d_{AC}-d_{BC})/2 \cdots\cdots\cdots\cdots (1a)$$

$$y=(d_{AB}+d_{BC}-d_{AC})/2 \cdots\cdots\cdots\cdots (1b)$$

而 D 至组合群 C 的距离 z 为，

$$z=(d_{BC}+d_{AC}-d_{AB})/2 \cdots\cdots\cdots\cdots (1c)$$

以上诸式中 $d_{AC}=(d_{A3}+d_{A4}+\cdots+d_{An})/(n-2)$

$d_{BC}=(d_{B3}+d_{B4}+\cdots+d_{Bn})/(n-2)$

(其中 d_{A3} 和 d_{B3} 表示除 A 和 B 外的第 3 个物种分别与 A 和 B 的距离，其他阿拉伯数字(在下标中)为该数字为序的物种分别与 A 和 B 的距离)

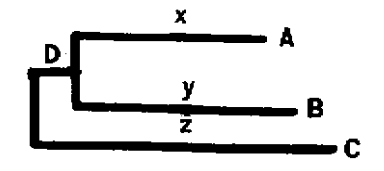


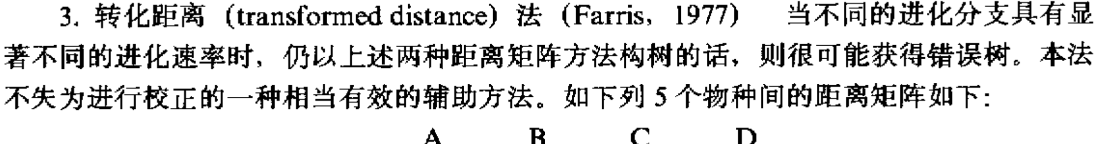
图 4 Fitch–Margoliash 方法的示意图

Fig. 4 Fitch–Margoliash method

然后将 A、B 组成一个新群 AB，列出新的矩阵。以类似方法将获得新的 x、y 和 z 值。依此继续下去直至将所有物种组成一棵最终树。

本法以观察值来构建进化树，基本上与 UPGMA 有类似的应用范围。计算机程序名为 EVOLVE。

3. 转化距离（transformed distance）法（Farris，1977） 当不同的进化分支具有显著不同的进化速率时，仍以上述两种距离矩阵方法构树的话，则很可能获得错误树。本法不失为进行校正的一种相当有效的辅助方法。如下列 5 个物种间的距离矩阵如下：



	A	B	C	D
B	18			
C	24	18		
D	20	14	6	
E	21	15	9	5

显然，若以 UPGMA 和 Fitch–Margoliash 法构树的话，则都获得错误树。若以另一物种（设想为与它们的共同始祖较接近的物种）作为参考系 r，r 至 A、B、C、D 和 E 的距离

分别为 16、10、12、8 和 9，应用公式，

$$d'_{ij} = (d_{ij} - d_{ir} - d_{jr}) / 2 + \bar{d}_r \cdots\cdots \quad (2)$$

将所有的 d_{ij} 转化成转化距离 d'_{ij}，式中 r 为参考系，$\bar{d}_r$ 为 r 至各物种的平均距离。当获得所有 d'_{ij} 后，即可用上述两种方法之一构建进化树。

李靖炎（1988）介绍的今祖法（present–day ancestor method）事实上是将 d_{ij} 转化，公式为 $d'_{ij} = d_{ij} - d_{ir} - d_{jr}$（$d'_{ij}$ 为转化后的距离）。实质上与转化距离法是等价的，同样是一种有效的辅助法，明确指出参考系可用接近共同始祖的今存物种来代替。

4. 邻接法（neighbor–joining method）（Saitou 等，1987） 首先算出星式树(starlike tree) 的分枝长度总和（图 5），

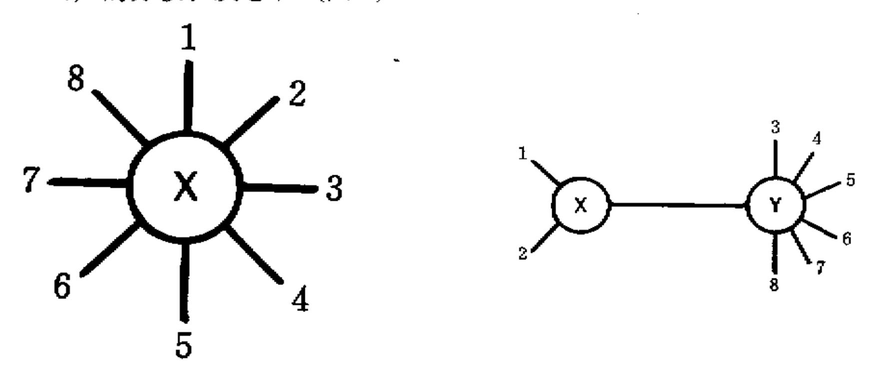


图 5 NJ(即邻接法)的星式树

Fig. 5 A starlide tree in NJ method

图 6 NJ 法中两内结 XY 分枝的构建

Fig. 6 A tree in which OTUs 1 and 2 are clustered in NJ method

$$S_0 = \sum_{i=1}^{N} L_{ix} = \frac{1}{N-1} \sum_{i<j} D_{ij} \cdots\cdots \quad (3)$$

（式中下标 i 和 j 指不同物种，X 为假设的共同祖先，N 为所分析的物种数目）接着找出物种间的最小距离如为 D_{12}，则可构成图 6。

X（物种 1 和 2 会聚点）至 Y（除物种 1 和 2 的其他诸物种会聚点）的分枝长度由下式求出，

$$L_{xy} = \frac{1}{2(N-2)} \left[\sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2)(L_{1x} + L_{2x}) - 2\sum_{k=1}^{N} L_{ky} \right] \cdots\cdots \quad (4)$$

式中 k 表示 1 和 2 物种外的其他物种；这样，物种 1 和 2 对其他物种的分枝长度总和为，

$$S_{12} = L_{xy} + (L_{1x} + L_{2x}) + \sum_{k=3}^{N} L_{ky}$$

$$= \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij} \cdots\cdots \quad (5)$$

然后可算出物种1和2至内结X的分枝长度，

$$L_{1x} = (D_{12} + D_{1z} - D_{2z}) / 2 \cdots\cdots \quad (6a)$$

$$L_{2x}=\ (D_{12}+D_{2z}-D_{1z})\ /2 \cdots\cdots\cdots\cdots\cdots\ (6b)$$

(上式中 $D_{1z}=\left(\sum_{k=3}^{N}D_{1k}\right)/\ (N-2)$ 和 $D_{2z}=\left(\sum_{k=3}^{N}D_{2k}\right)/\ (N-2)$，下标 z 为物种 1 和 2 外的其他物种)；继后将物种 1 和 2 合并成新群，甚至其他物种的距离可由公式 $D_{(1-2)k}=\ (D_{1k}+D_{2k})\ /2$ 求出。仿上式继续运算，直至形成一棵最终树。

Saitou 与 Nei 经数学论证，由邻接法所获得的进化树是替代总数为最小的树；他们还通过计算机模拟，指出此法优于几乎所有的其他距离矩阵法。

三、最大似然法

以各种假设的进化数学模型对观察数据进行检验，选出具有最大似然函数的模型构树，以解释进化过程。

1. Langley 与 Fitch (1974) 法　该模型假设 $v_i=\lambda t_i$，其中基因替代率 λ 按 Poisson 分布。若有 4 个物种的一进化树，经最大简约法或距离矩阵法获得诸 x_i 值后，即可以下述公式求出最大似然函数值 (图 7)，

$$L=e^{-v_1}\frac{v_1^{x_1}}{x_1!}\cdot e^{-v_1}\frac{v_1^{x_2}}{x_2!}\cdot e^{-(v_1+v_2)}\frac{(v_1+v_2)^{x_3}}{x_3!}\cdot e^{-(v_1+v_2)}\frac{(v_1+v_2)^{x_4}}{x_4!}$$

以偏导数法可求出 v_i 的估计值 $\hat{v}_i$，然后以公式，

$$\chi^2=\sum_{i=1}^{m}\frac{(x_i-\hat{v}_i)^2}{\hat{v}_i}\cdots\cdots\ (8)$$

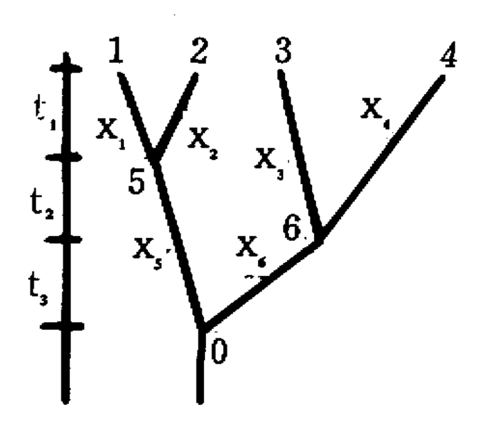


图 7　Langley-Fitch 方法的示意图
(左为时间尺度，右为进化树)
Fig. 7　The langley-Fitch method (A time scales in left; a molecular tree in right)

检验观察值和期望值的吻合度，以判断基因替代率在进化历程中是否符合 Poisson 分布 (x^2 的自由度为 $m-p$，m 为观察值的数目，p 为 $\hat{v}_i$ 的数目)。

2. Felsenstein (1981) 法 (PHYLIP 程序)　如仍以图 7 为例，Felsenstein 用下式计算似然函数，

$$L=\sum_{s_0}g_{s_0}\left[\sum_{s_5}p_{s_0s_5}(v_5)p_{s_5s_1}(v_1)p_{s_5s_2}(v_1)\right]\cdot\left[\sum_{s_6}p_{s_0s_6}(v_2)p_{s_6s_3}(v_3)p_{s_6s_4}(v_4)\right]\cdots\cdots\cdots\ (9)$$

对任何 $p_{ij}(v)$ 来说，Felsenstein 提出下述随机替代模型，

$$p_{ij}(v)=e^{-v}\delta_{ij}+\ (1-e^{-v})\ g_j\cdots\cdots\cdots\ (10)$$

〔上两式中，g_{s_0} 为图 7 中根 0 状态（即取 A、C、G 或 T）时的前概率，$p_{sisj}(v)$ 为分枝长度为 v 时从 i 状态变化至 j 状态的转移概率。当 $i=j$ 时 $\delta_{ij}=1$，而当 $i\neq j$ 时 $\delta_{ij}=0$，δ_j 为第 j 种碱基代替后的固定概率〕。实际运算时，本法需化费大量计算机运算时间，但其优点是需要的假设很少，因而更符合正确的进化树。

3. Lake (1987) 的进化简约法　此法又称不变量 (invariant) 法。基本思路为：在进化过程中，转换和颠换有着根本不同的生物学意义，从而排除了平行进化和返祖进化对构建单系类群进化树的干扰作用（吕宝忠，1991a）。

对一个含 4 个物种的无根树来说，其 4 条分子序列中的每一同源位点核苷酸理论上都可构成 256 种不同构型的四聚体，但只有其中 96 种构型具有信息，这 96 种构型又可归并成 xxzz，xyzw，xyzz，xxzw，xzxz，xzyw，xzyz，xzxw，xzzx，xzwy，xzzy 和 xzwx 共 12 种具有信息的构型，此外构型均无信息（x 代表任何一种碱基，y 代表与之有转换关系的碱基，z 代表与之有颠换关系的碱基，w 代表与 z 有转换关系的碱基）。如果 $P_{(xxzz)}+P_{(xyzw)}-(P_{(xyzz)}+P_{(xxzw)})$ 为正值而上述第 5 至第 8 构型的相应类似式以及上述第 9 至第 12 构型的类似式接近 0，则第 1 种树即 ${}^{1}_{2}\succ\!\!-\!\!\prec{}^{3}_{4}$（1、2、3 和 4 代表 4 种不同物种）为历史上曾出现过的进化树的可能性最大。然而如此判断毕竟太粗，于是 Lake 提出以 $\chi^2=\dfrac{(P-B)^2}{P+B}$ 检验公式（其中 P 项为 $P_{(xxzz)}+P_{(xyzw)}$ 或 $P_{(xzxz)}+P_{(xzyw)}$，或 $P_{(xzzx)}+P_{(xzwy)}$，它们分别代表第 1 种树或第 2 种树即 ${}^{1}_{3}\succ\!\!-\!\!\prec{}^{2}_{4}$ 或第 3 种树即 ${}^{1}_{4}\succ\!\!-\!\!\prec{}^{2}_{3}$；B 项为第 1、2 或 3 种树的背景项，分别为 $P_{(xyzz)}+P_{(xxzw)}$、$P_{(xzyz)}+P_{(xzxw)}$ 或 $P_{(xzzy)}+P_{(xzwx)}$；χ^2 的自由度为 1）。当仅上述某一种树的 $\chi^2>3.84$，而其他二种树的 χ^2 值 <3.84 时，则 >3.84 的树即为所需要的树。

应用本法 Lake 成功地定量描述了线粒体和叶绿体的内共生起源，正确地解决了原核类和真核类的单系起源，雄辩地说明该法具有的强大生命力。

Sidow 等 (1990) 提出了一种新的构树法——组合统计(compositional statistics)法，自认为是本法的延伸。

晚近 Lake (1991) 认为物种配对时的先后次序对进化简约法构建的最终树有影响，看来尚待改进。

Y. -X. Fu 与 W. -H. Li (1992) 从半群理论论证了本法，为该法提供严格的数学证明。

两类不同的重抽样技术

构建分子进化树的分子数据，包括蛋白电泳数据、内切酶图谱数据、DNA 杂交数据以及分子序列数据，可能除序列数据外都存在较大的抽样误差。重抽样技术可校正抽样误差。晚近 Marshall (1991) 已用自助法对 DNA 杂交数据作出区间估计。另一类重抽样技术虽已有人 (Mueller 等，1982) 用来估计遗传距离的变异性（该技术为折刀法），但尚未见用于分子数据。

分子进化树构建的展望

本文所述三大类构树法中，距离矩阵法应用相当普遍，其算法基础建立在分子钟假设上，应用在有较近亲缘关系的物种或群体的分化研究上能获得满意结果，但难以排除平行进化等的干扰；最大简约法并不需要分子钟假设，其算法基础建立在所研究的进化历程应符合最小替代数原则上，因而也可用于较远亲缘关系的物种分化研究，但也不能排除平行进化等的干扰；最大似然法建立在比较严密的概率统计基础上，所需假设最少，尽管运算时间长且还未臻成熟，但前景肯定最为乐观。

当前分子进化树的构建绝大多数以生物大分子的一级结构数据为材料，虽已获得相当可喜的进展，但应清楚地看到，生物大分子的功能主要体现在其三维结构上（刘次全等，1990）。事实上，三维结构的进化应该是分子进化树构建的最可靠依据。吕宝忠等（1991 b）曾初步进行了这类研究。但由于这类数据需依赖X光衍射、圆二色性、电子显微镜以及核磁共振等高新技术，有一个积累过程，更由于这类数据的非线性系统，尚需大力发展这方面的数学方法，但毫无疑问，客观地再现历史上曾发生过的如此宏伟的进化历程，必须向这个新的更高层次冲刺！

参 考 文 献

吕宝忠. 1991a. 进化构树法、分支理论及选择／中性学说. 自然杂志，**14**: 734—738.

吕宝忠、陈捷，顾健人. 1991b. 从分子进化角度探讨P^{21}高级结构变异与癌变的关系. 遗传，**13** (1) 41—43.

吕宝忠，陈捷. 1992. 进化距离与相对替代率的比较研究. 遗传学报，**19**: 397—402.

刘次全，黄京飞，王莹. 1990. 分子进化研究中的一些问题. 动物学研究，**11** (2): 167—172.

李靖炎. 1988. 分子进化研究中的今祖法，其理论基础、存在的问题和解释. 动物学研究，**9** (2): 141—150.

Beintema, J. J., Fitch, W. M. & Carsana, A. 1986. Molecular evolution of pancreatic-type ribonucleases. *Mol. Biol. Evol.*, **3**: 262–275.

Farris, J, S. 1977. On the phenetic approach to vertebrate classification. In M. K. Hecht *et al.* eds., Major Patterns in Vertebrate Evolution, Plenum Press, New York. 823–850.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**: 368–376.

Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. *Amer. Natur.*, **3**: 223–257.

Fitch, W. M. & Margoliash, E. 1967. Construction of phylogenetic trees. *Science*. **155**: 279–284.

Fu, Y. -X. & Li, W. -H. 1992. Necessary and sufficient conditions for the existence of linear invariants in phylogenetic inference. *Math. Biosci.*, **108**: 203–218.

Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences; evolutionary parsimony. *Mol. Biol. Evol.* **4**: 167–191.

Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.*, **8**: 378–385.

Langley, C. H. & Fitch, W. M. 1974. An examination of the constancy of the rate of molecular evolution

. *J. Mol. Evol.* 3: 167–177.

Marshall, C. R. 1991. Statistical tests and boostrapping: assessing the reliability of phylogenies based on distance data. *Mol. Biol. Evol.*, 8: 386–391.

Mueller, L. D. & Ayala, F. J. 1982. Estimation and interpretation of genetic distance in empirical studies. *Genet. Res. Camb.*, 40: 127–137.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4: 406–425.

Sidow, A. & Wilson, A. C. 1990. Compositional statistics: an improvement of evolutionary parsimony and its application to deep branches in the tree of life. *J. Mol. Evol.*, 33: 51–68.

* *

增 刊 预 告

我刊计划于今年 9 月上旬出版第 14 卷第 3 期增刊——理论生物学研究论文专辑。

"理论生物学"是一门以理论物理学的原理为指导，以计算科学的方法处理生物学问题的极其重要的交叉学科。

本专辑将较集中地反映我国科研工作者在量子生物学、理论进化生物学、计算分子生物学和生物大分子结构多样性等方面的理论研究情况，突出新理论、新设想以及运用多学科综合手段研究所取得的结果，有些工作在理论计算与生物学问题的紧密结合上颇具特色。

专辑包含生物系统中的非对称性；生物系统中的有序性；Eucarytic DNA Methylation and Gene Mutation; STM Barrie-height Images and Topgraphic Images of Denatured DNA；三碱基结合体模型分析；三碱基结合体的能力学分析；λ 噬菌体 DNA 核苷酸序列可能生成三链辫状结构的理论研究；辫状三链 DNA 在纳米尺度上的精细结构；蛋白质与遗传信息流；三链 DNA 的理论研究等 20 余篇论文。

希望增刊的出版能对广大读者有所帮助。欢迎订阅。

《动物学研究》编辑部

1993 年 3 月